CRICKET MATCH ANALYSIS AND PREDICTION

Abhishek Thamatam (TF53379)

Sai Ruthvik Anantapalli (CW54735)

Rohith Challa (ZI04298)

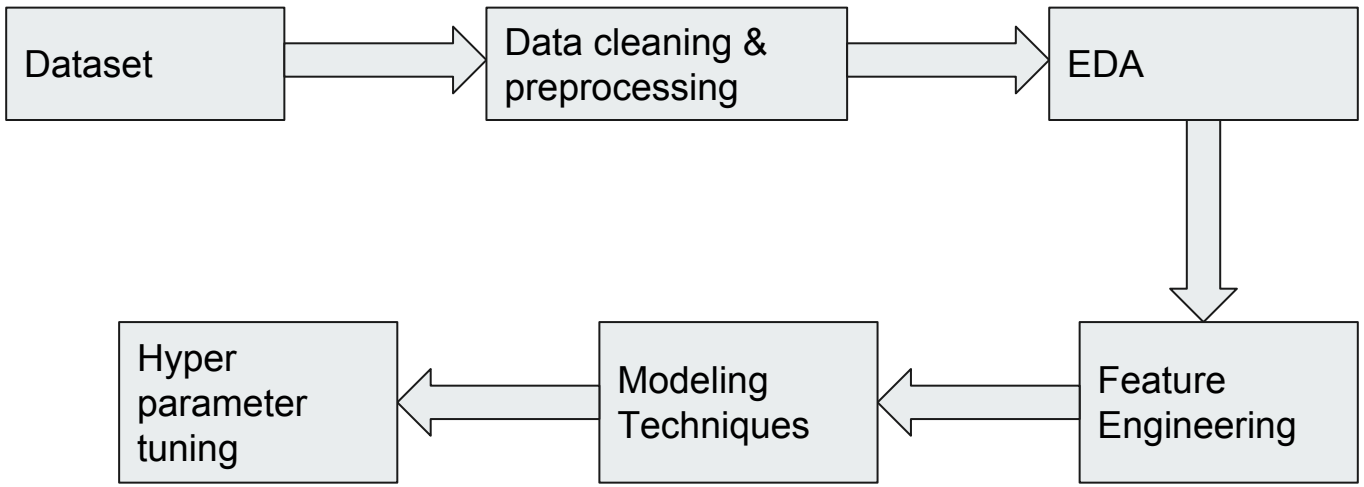
CONTENTS

INTRODUCTION

- The main goal of this project is to make use of Spark functionalities to analyse and predict the result of a cricket match.
- As part of our project we :

  1. cleaned the data
  2. conducted exploratory data analysis
  3. developed and trained machine learning models and
  4. made predictions

PROJECT FLOW



DATASET OVERVIEW

- The dataset we used in this project is from Kaggle and contains information on team names, innings, scores, venue, home and away teams, match winner etc.
- This dataset consists of 1000 rows and 45 columns.
- Few of our potential features include :

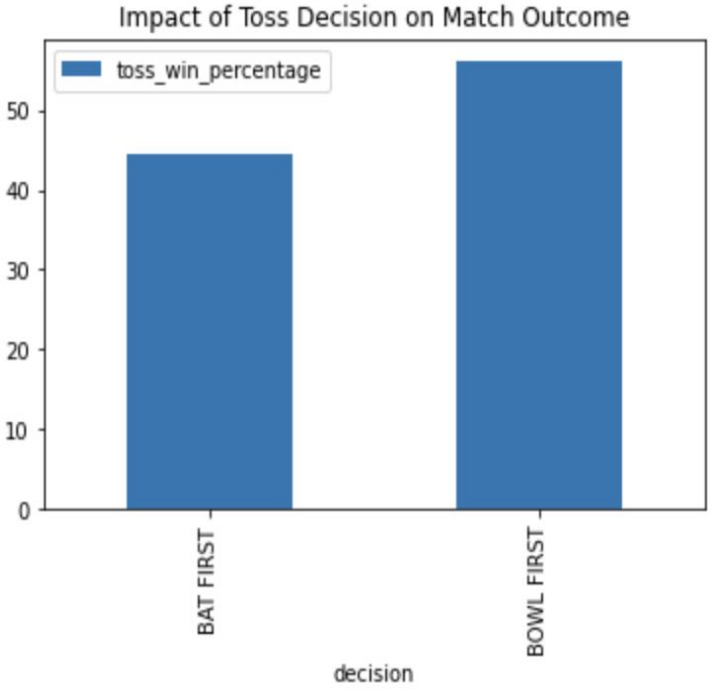
  1. home and away teams
  2. venue
  3. toss decision
  4. innings wise scores

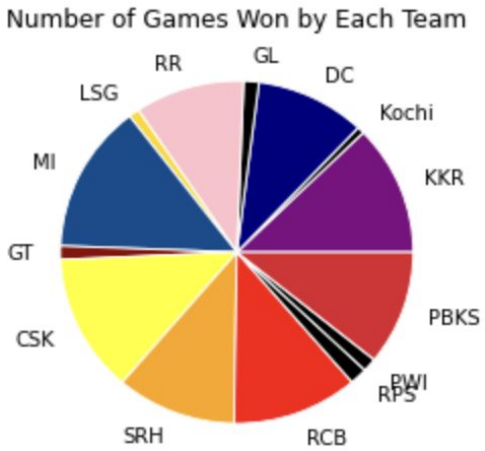
- The target variable in this project is ‘is_home_team_winner’

DATA CLEANING AND PRE-PROCESSING

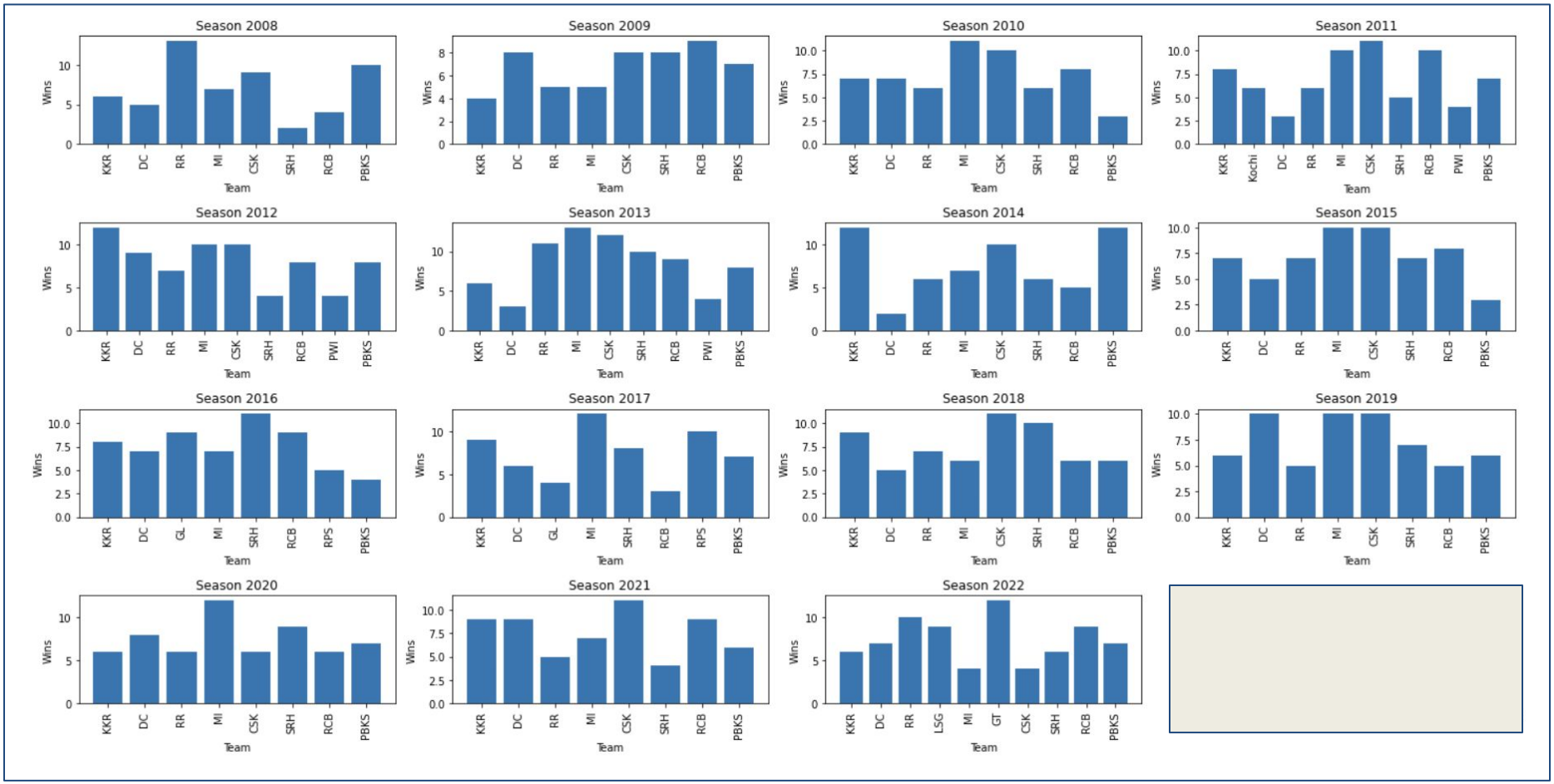
- The dataset required cleaning, including handling missing values, identifying and correcting errors, and eliminating the inconsistencies.
- For data pre-processing we transformed the existing features into features which is more useful for analysis and prediction.

DATA ANALYSIS AND VISUALIZATION

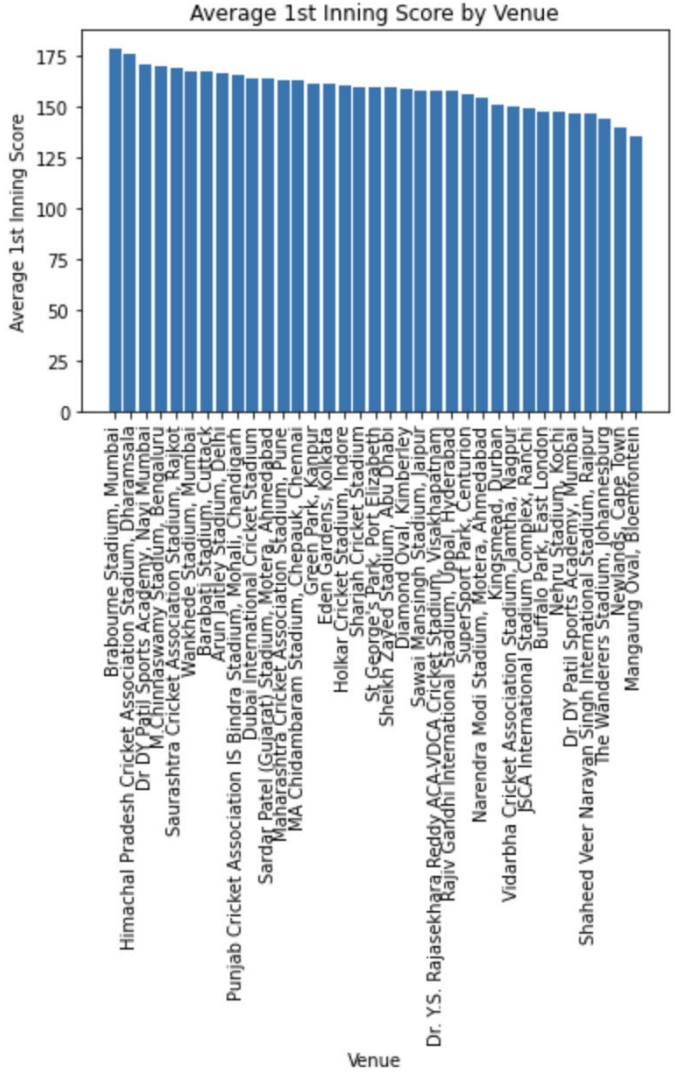




DATA ANALYSIS AND VISUALIZATION



DATA ANALYSIS AND VISUALIZATION



MACHINE LEARNING

- Apache Spark is a powerful distributed computing system that can also be used for machine learning tasks.
- Spark's MLlib library offers a wide range of machine learning algorithms for classification, regression, clustering, and more.

FEATURE ENGINEERING

Feature engineering is the process of transforming raw data into features that can better represent the underlying problem to the predictive models. In this project, we have performed several feature engineering techniques such as:

- Encoding categorical variables('home_team', 'away_team', 'toss_won', 'venue', 'home_captain', etc)
- Creating new features (home_run_rate, away_run_rate, home_economy_rate, away_economy_rate, score_difference, wicket_difference)

Modeling Techniques

This is classification problem.

We have used several modeling techniques to predict the outcome of cricket matches, including:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- Support Vector Classifier

Model Evaluation Metrics

To evaluate the performance of our predictive models, we have used several evaluation metrics such as:

- Accuracy
- Precision
- Recall
- F1-score

Results

- Logistic Regression:

```
Accuracy: 0.937269
Precision: 0.938799
Recall: 0.937269
F1-score: 0.936971
```

- Gradient Boosting:

```
Accuracy: 0.97417
Precision: 0.974344
Recall: 0.97417
F1-score: 0.974135
```

- SVC:

```
Accuracy: 0.9631
Precision: 0.963392
Recall: 0.9631
F1-score: 0.963032
```

- Random Forest:

```
Accuracy: 0.98893
Precision: 0.989147
Recall: 0.98893
F1-score: 0.988915
```

CONCLUSION

In this project, we have analyzed cricket match data and used machine learning techniques to predict the outcome of cricket matches. We have obtained valuable insights that can be useful for cricket teams and Fantasy sports industry. Our models have achieved good performance, but there is still room for improvement. Overall, this project demonstrates the potential of machine learning techniques in analyzing and predicting the outcome of cricket matches.

Future Work

- Incorporating more data such as player performance statistics and weather data
- Using dashboards(Tableau/PowerBI etc)
- Exploring more advanced machine learning models such as neural networks and ensemble models
- Live prediction

Thank you!!